INHerit-SG: Incremental Hierarchical Semantic Scene Graphs with RAG-Style Retrieval

Author Names Omitted for Anonymous Review. Paper-ID [643]

***Abstract*—Driven by advancements in foundation models, semantic scene graphs have emerged as as a prominent paradigm for high-level 3D environmental abstraction in robot navigation. However, existing approaches are fundamentally misaligned with the needs of embodied tasks. As they rely on either offline batch processing or implicit feature embeddings, the maps can hardly support interpretable human-intent reasoning in complex environments. To address these limitations, we present INHerit-SG. We redefine the map as a structured, RAG-ready knowledge base where natural-language descriptions are introduced as explicit semantic anchors to better align with human intent. An asynchronous dual-process architecture, together with a Floor-Room-Area-Object hierachy, decouples geometric segmentation from time-consuming semantic reasoning. An event-triggered map update mechanism reorganizes the graph only when meaningful semantic events occur. This strategy enables our graph to maintain long-term consistency with relatively low computational overhead. For retrieval, we deploy multi-role Large Language Models (LLMs) to decompose queries into atomic constraints and handle logical negations, and employ a hard-to-soft filtering strategy to ensure robust reasoning. This explicit interpretability improves the success rate and reliability of complex retrievals, enabling the system to adapt to a broader spectrum of human interaction tasks. We evaluate INHerit-SG on a newly constructed dataset, HM3DSem-SQR, and in real-world environments. Experiments demonstrate that our system achieves state-of-the-art performance on complex queries, and reveal its scalability for downstream navigation tasks.**

I. Introduction

The focus of robotic mapping has been steadily evolving. Traditionally, robots prioritized high-precision metric reconstruction to ensure safe navigation [33, 10, 4]. However, the rise of embodied AI is shifting this focus toward semantic interaction. An agent operating in human environments must understand vague, language-driven instructions rather than just coordinate goals. In this context, strict metric localization is not a necessity in many modern embodied tasks. Benchmarks such as Object Goal Navigation (ObjectNav) and Vision Language Navigation (VLN) [37, 41, 32] consider an episode successful if the agent stops within a 1-meter radius of the target. This reflects a shift from geometric accuracy to semantic understanding, which is sufficient for the robot to find and interact with the object. A robot does not require a perfect point cloud to locate and identify a cup. Instead, it requires a semantically meaningful index to bridge the gap between human language and physical space.

We argue that, to effectively serve embodied intelligence tasks, the mapping system for robots need to satisfy several essential requirements. **Structured.** The map should organize the environment into a multi-level topology rather than a flat collection of features, mirroring human spatial cognition to support scalable reasoning. **Semantically Rich.** The map must contain deep visual and semantic attributes. This is essential for grounding abstract human concepts into concrete physical entities. **On-the-fly.** The system should support incremental maintenance during exploration. While strict real-time synchronization is unnecessary, the map must capture meaningful semantic changes during exploration without relying on heavy offline post-processing. **Interpretable.** The retrieval mechanism must go beyond opaque embedding matching. It requires robust reasoning capabilities to accurately parse complex language constraints and ensure verifiable results.

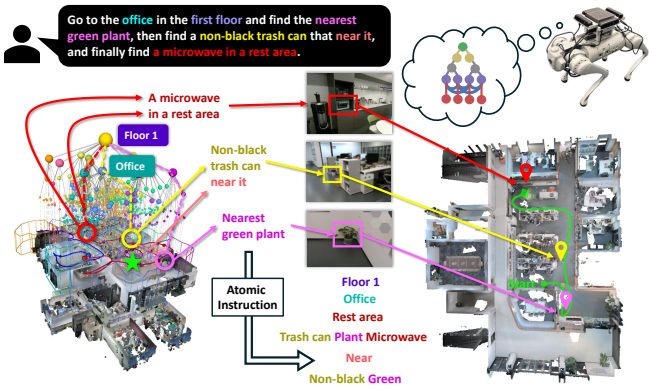

Fig. 1. **INHerit-SG Overview.** Our system build a hierarchical semantic memory during online exploration and operate closed-loop retrieval. (Left) The hierarchical scene graph of a real-world office building built through incremental mapping. (Right) The robot parses a complex query into structural constraints and follows the retrieval pipeline to complete the task sequentially.

But existing methods struggle to satisfy all these requirements simultaneously. Recent 3D semantic mapping has evolved along two main axes, flat open-vocabulary feature fields and structured hierarchical scene graphs. While flat representations [12, 16] perform well in zero-shot recognition, they encode maps as dense, point-aligned embeddings without explicit multi-level structure. As a result, flat feature-field representations are neither structurally expressive nor interpretable, making it difficult to support scalable reasoning over complex spatial and semantic constraints. Meanwhile, current structured methods [17, 44, 26] provide richer geometric and topological details but often incur high computational costs and storage redundancy. Some real-time systems, such as Hydra [17], incorporate more explicit segmentation labels and geometric descriptors for storage. However, such geometric descriptors and categorical labels still lack semantic richness

and expressiveness required to ground abstract human intent.

In parallel, retrieval mechanisms in current embodied systems [45, 39] typically operate in an open-loop manner, relying primarily on embedding similarity for recall. This strategy is fragile to complex logical structures such as negation or chained spatial relations and frequently produces false positives without explicit verification. Although recent navigation-focused methods [47, 36, 8, 51, 49, 46] have begun to integrate confidence calibration, graph prompting, or active exploration, most still lack a systematic closed-loop verification mechanism to audit candidates against full semantic intent. As a result, current semantic mapping pipelines remain poorly aligned with the logical reasoning demands of embodied interaction, particularly in terms of interpretability.

To achieve these requirements, we propose **INHerit-SG**, a lightweight scene graph system designed for long-term embodied execution. We argue that visual features alone are insufficient for representing semantics. Natural language, by contrast, is explicit, compositional, and aligned with human understanding. Therefore, beyond image features, we store natural-language descriptions in the map as **Semantically Rich** representation grounded in human concepts. We redefine the map as a **Structured**, RAG-ready knowledge base organized into a multi-level Floor–Room–Area–Object hierarchy, where visual features provide perceptual grounding and natural-language descriptions serve as explicit semantic anchors. For **On-the-Fly** efficiency, our system employs a event-triggered mechanism that updates topology only upon meaningful semantic changes Furthermore, we couple this mapping engine with an **Interpretable** closed-loop retrieval pipeline. This system moves beyond opaque embedding matching by utilizing multi-role LLM parsing for logical constraint enforcement and VLM-based visual auditing, ensuring precise adherence to complex user intents.

In summary, we make the following contributions:

1) We propose INHerit-SG, a hierarchical scene graph framework that organizes the map as a RAG-style, language-indexed knowledge base. By retaining visual features for perceptual grounding while treating natural language as the semantic anchor, the map becomes directly compatible with human reasoning and complex queries.
2) We design an asynchronous dual-process architecture with an event-triggered update mechanism. INHerit-SG decouples geometric segmentation from semantic reasoning and reorganizes the graph only when meaningful semantic events occur, enabling scalable, incremental mapping.
3) We develop an interpretable closed-loop retrieval pipeline that enforces logical constraints through LLM parsing and VLM-based verification, significantly improving reliability for complex queries beyond similarity-based retrieval.
4) We construct HM3DSem-SQR, a dataset to test high-level reasoning and fine-grained retrieval, including logical negations, spatial relationships, and complex attribute constraints. *Source code and dataset will be released to benefit the community.*

II. Related Work

A. Open-Vocabulary Semantic Mapping

The integration of Vision-Language Models (VLMs) has fundamentally shifted semantic mapping from closed-set label classification to open-vocabulary understanding. Early approaches in this domain leveraged foundation models to construct dense, semantic feature fields. Methods such as ConceptGraphs [12], VLMaps [16], OpenScene [31], LERF [21], and OpenMask3D [42] project high-dimensional features directly into 3D space. Recent advancements including Open3DIS [29], FMGS [52], SplatSearch [28], OVIR-3D [25], and OmniMap [6] have further refined this paradigm through instance segmentation and Gaussian Splatting integration. While these flat representations excel at zero-shot recognition, they typically organize the map as dense collections of point-aligned or voxel-wise embeddings. Although effective for simple queries, they generally lack explicit hierarchical abstractions, which can lead to poor scaling in large environments and poor efficiency for complex spatial queries.

To enable deeper spatial reasoning, researchers have developed structured 3D scene graphs. Offline methods like Open3DSG [22], HOV-SG [44], FSR-VLN [50], SceneGraphLoc [27], and OpenIN [43] construct rich hierarchies enabling relationship modeling. However, these approaches typically rely on heavy global optimization or batch processing, limited in online applicability. Planning frameworks like SayPlan [35] circumvent this by assuming pre-constructed graphs. Conversely, real-time systems such as Hydra [17], Clio [26], Describe Anything [11], ZING-3D [38], and The Bare Necessities [20] focus on incremental construction. Despite their efficiency, several of these systems still largely rely on high-dimensional embeddings or relatively simple categorical tags, which can limit fine-grained interpretability and compositional reasoning. While most open-vocabulary methods use latent embeddings as the primary semantic representation, some real-time systems (e.g., Hydra) incorporate more explicit geometric descriptors and segmentation labels. However, these approaches are still limiting interpretability and are weak for reliable language-grounded reasoning.

B. Incremental Updates and Global Consistency

For long-term autonomy, a map must be a living entity capable of adapting to dynamic changes. Approaches such as DualMap [19] and Khronos [40] address this by maintaining spatio-temporal consistency through hybrid abstract-concrete layers or unified metric-semantic formulations. Similarly, works like DynamicGSG [9], REACT [30], and MoMa-LLM [15] focus on real-time attribute clustering and updating object geometry to handle object dynamics. Additionally, methods including GraphPad [1], EmbodiedRAG [3], and RoboEXP [18] emphasize inference-time updates or exploration-driven graph expansion. Despite these advances, many update policies are still primarily triggered by geometric changes or fixed time intervals. While some recent works begin to incorporate object-level or semantic change detection, fully semantically-aware topological event triggering remains an open challenge.

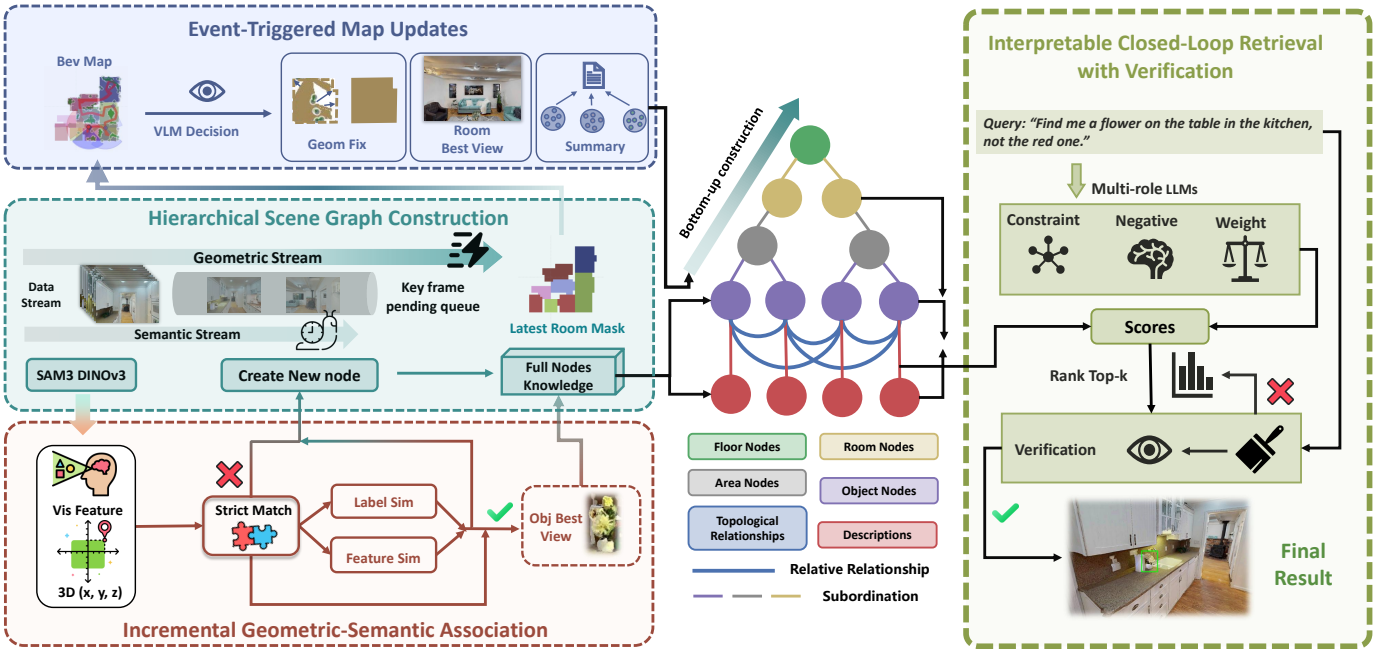

Fig. 2. **The INHerit-SG Framework.** The system bridges real-time mapping with logic-aware retrieval. **(Left)** The pipeline employs a dual-stream architecture to balance tracking and reasoning. A *Event-Triggered Map* module (top-left) optimizes topological updates based on VLM decisions, while the *Incremental Association* block (bottom-left) fuses SAM3/DINOv3 features to instantiate nodes. **(Center)** The resulting data structure is a multi-level scene graph that explicitly models topological relationships. **(Right)** Complex queries are decomposed by *Multi-role LLMs* into specific constraints, including negation and weights. The system ranks candidates using a scoring function and executes a final VLM *Verification* step to ensure precise intent grounding.

C. Semantic Retrieval and Verification

The utility of a semantic map is ultimately defined by how accurately a robot can retrieve objects from it. Inspired by Retrieval-Augmented Generation (RAG) in NLP [24, 7, 14], embodied retrieval systems typically map natural language queries directly to map embeddings. Methods such as Embodied-RAG [45], GraphEQA [39], LLM-Grounder [46], and RAG-3DSG [5] perform top-k recall based on vector similarity. Specialized variants [23, 48, 13] extend this to affordance-aware and ontology-based memory. These approaches frequently struggle with logical structures where visually similar objects may be incorrectly prioritized. Recent works have sought to mitigate these issues by integrating retrieval with active exploration, confidence calibration, and graph prompting. Explore until Confident [36] uses conformal prediction for uncertainty-aware stopping. LLM-Grounder [46] performs explicit relation evaluation. Approaches like SG-Nav [47], Explore until Confident [36], RoboHop [8], and MTU3D [51] combine graph prompting with navigation. To handle temporal context, Mem2Ego [49] and ReMEmbR [2] align global memory with ego-centric cues. Despite these advances, existing systems generally lack an explicit closed-loop verification mechanism to audit retrieved candidates against the logical intent of the query, leaving them vulnerable to false positives in cluttered or complex environments.

III. Technical Approach

We propose **INHerit-SG**, a unified framework for incremental hierarchical semantic scene graph construction and closed-loop retrieval. Our approach is designed around two core principles: (1) *Geometric Stability for Semantic Anchoring*, ensuring that high-level semantics are grounded in a robust geometric skeleton; and (2) *Interpretable Verification*, shifting from black-box similarity matching to a transparent, logic-driven retrieval pipeline.

As shown in Figure 2, our system processes a stream of RGB-images and camera poses to maintain a dynamic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The process begins with the Hierarchical Construction Module (Sec. III-A). Here, a fast geometric stream builds structural layers, including *Floors* (L_0) and *Rooms* (L_1), while a semantic stream instantiates atomic *Objects* (L_3). Next, the Incremental Association Module (Sec. III-B) fuses temporal observations while preventing redundancy during tracking. The Map-Conditioned Update Module (Sec. III-C) generates intermediate *Functional Areas* (L_2). This module refreshes the graph topology only when significant semantic events occur. Finally, the Closed-Loop Retrieval Module (Sec. III-D) handles user interaction. It parses instructions into structural constraints and performs a visual audit via a VLM to output a verified 3D target location.

A. Hierarchical Scene Graph Construction

The semantic memory is built upon a robust geometric foundation. We employ a hierarchical construction strategy distributed across the dual-stream architecture to balance mapping accuracy with computational efficiency. Importantly, node representations are designed from the outset to align with RAG-style knowledge organization, allowing the map to

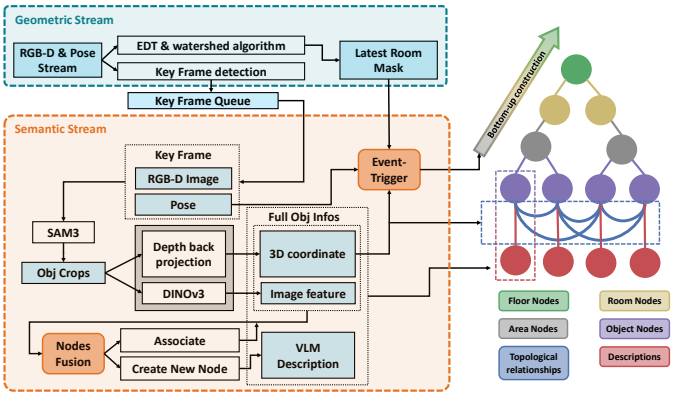

Fig. 3. **Dual-Stream Construction Pipeline.** We decouple mapping into a *Geometric Stream* (top) for online room segmentation and an asynchronous *Semantic Stream* (bottom) for fine-grained object reasoning. These threads converge via an *Event-Trigger* mechanism, which incrementally construct the hierarchical scene graph from the bottom up.

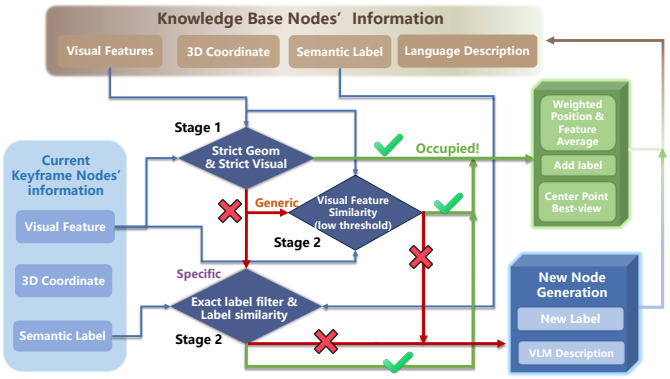

Fig. 4. **Incremental Node Association Logic.** The association process follows a two-stage cascade. **Stage 1** filters high-confidence matches using strict geometric and visual thresholds. **Stage 2** resolves ambiguities based on semantic specificity, enforcing label consistency for known categories while relying on high visual similarity for generic, open-vocabulary objects.

function directly as a structured, queryable knowledge base.

Geometric Stream: Dense Topology & Keyframe Gating (L_0, L_1). As illustrated in Fig. 2, the *Geometric* steam acts as the backbone for structural stability. It continuously integrates the dense RGB-D stream into a voxel-based occupancy map. We perform room segmentation (L_1) directly on this accumulated free space using a Euclidean Distance Transform (EDT) and watershed algorithm. Simultaneously, Vertical motion is monitored to instantiate Floor nodes (L_0), enabling automatic structural separation across floors.

Besides, we implement a visual gating mechanism to regulate data flow to the semantic stream. We extract global DINOv3 features and calculate cosine similarity against the last processed frame. When this similarity drops below a threshold, the system pushes the frame to the Semantic Queue with its floor ID. This queue serves as a buffer, holding selected keyframes for asynchronous, fine-grained analysis by the *Semantic* semantic stream. This ensures that semantic reasoning operates only on informative keyframes while geometric tracking remains lightweight and continuous ($\approx$ 2Hz).

Semantic Stream: Object Instantiation (L_3). The *Semantic* thread operates asynchronously on the Semantic Queue to instantiate fine-grained object nodes (L_3). For each keyframe, we use the Segment Anything Model (SAM3) to generate instance masks and back-project their centroids into 3D coordinates. To mitigate the temporal latency inherent to this heavy inference, we implement a floor-consistent asynchronous query strategy. Rather than relying on the occupancy state synchronous with the keyframe timestamp, the thread queries the Geometric Stream for the latest accumulated Room Segmentation Mask associated with the keyframe's specific Floor ID. Since the Geometric Stream continuously integrates dense topological data, this retrieved mask offers superior boundary completeness and segmentation accuracy compared to the partial state available at the time of capture. This ensures that objects from previous keyframes are registered within the most comprehensive geometric layout available, guaranteeing robust room assignment regardless of the robot's subsequent navigation across different rooms or floors.

RAG-Oriented Lightweight Node Representation. A key design choice in INHerit-SG is to treat the scene graph as a lightweight, RAG-aligned knowledge base rather than a geometry-heavy map. Departing from traditional embedding-heavy metric maps, we explicitly decouple semantic storage from geometric reconstruction to ensure interpretability and scalability. We adopt a compact, reference-based storage strategy where Object nodes (L3) host metadata, including semantic tags, visual embeddings, and 3D centroids with a reference pointer to their optimal observation keyframe. The raw high-resolution imagery is managed in a separate global hash table. This design establishes a memory-efficient many-to-one mapping between objects and keyframes, as multiple objects often share the same best-view perspective. During verification, the system dynamically accesses the specific best-view image via this index. Higher-level Area (L_2) and Room (L_1) nodes aggregate context via IDs and semantic summaries, with Room nodes additionally preserving 2D segmentation masks for topological grounding. The global structure is serialized efficiently via directed graphs and structured tables. Compared to volumetric or pointcloud maps, this design drastically reduces memory usage while making the graph directly compatible with language-driven retrieval. We quantitatively validate this significant advantage in Section IV.

B. Incremental Geometric-Semantic Association

Merging new observations into stable graph nodes is critical for preventing semantic drift and redundancy. Rather than relying on offline global optimization, INHerit-SG resolves data association incrementally, ensuring that nodes remain stable while accommodating both known categories and open-vocabulary objects.

Open-Vocabulary Association Logic. A naive spatial or visual matching strategy easily leads to over-merging in open-vocabulary settings. Therefore, we design a two-stage fusing cascade as illustrated in Fig. 4. First, a strict geometric gate

associates observations that have high spatial overlap and high visual similarity with existing nodes. Second, for ambiguous cases, the system decides upon semantic specificity. For objects with specific labels, we enforce strict label consistency while relaxing spatial constraints. For open-vocabulary objects outside predefined categories, we retain them and associate instances using a high visual-similarity threshold. This ensures that the system remains compatible with open-set environments, preventing *generalization pollution* where visually distinct but spatially adjacent unknown objects are erroneously merged. Upon a successful merge, we execute a Best-View Update. The system compares the bounding boxes of the current observation and the existing node. We retain the keyframe path where the object's bounding box is closer to the image center, ensuring that the node is always linked with the most informative visual perspective.

Local Spatial Topology Construction. Following the update of object nodes in the current frame, the system establishes spatial edges between L_3 Room nodes to support relational reasoning. We first apply a distance-based clustering on the horizontal plane to identify spatially adjacent groups within the current view. Within each cluster, pairwise relationships are inferred using a configurable hybrid submodule. The system either adopts a Geometric Mode that calculates heuristics based on 3D bounding box offsets and vertical proximity, or uses a VLM Mode to analyze the annotated RGB image to deduce complex semantic relations. These validated edges are inserted into the global spatial graph, enabling the system to effectively resolve spatial-relational queries.

C. Event-Triggered Map Updates

A key question in incremental semantic mapping is not how to update the map, but when the map should be reorganized. Rather than relying on time or motion as triggers, our system treats semantic topology changes as the primary signal for reorganization. As is illustrated in Fig 5, we propose a Event-Triggered mechanism that mimics the marginalization process in SLAM [10, 33, 4], triggering high-level summarization only when the topological belief stabilizes.

We first employ a supervisor module that intelligently monitors the robot's exploration state to trigger updates. A Hard Trigger is activated by discrete state changes, such as floor switches. A Soft Trigger is designed using a novel VLM-based decision-making process. We frame the VLM as a high-level supervisor, providing it with a task-specific, dynamically generated Bird's-Eye View (BEV) map. As shown in Fig. 5, the BEV visualizes key topological data: room segmentation masks (colored overlays), the current trajectory (red line), and historical update points (blue wedges). Crucially, the blue wedges fade over time, providing a visual cue for temporal staleness. The VLM analyzes this map to detect significant events, such as entering a New Area or completing a loop closure. The system triggers an update only if the VLM confirms that the topological change warrants a global refresh. This ensures we do not waste resources on redundant motion.

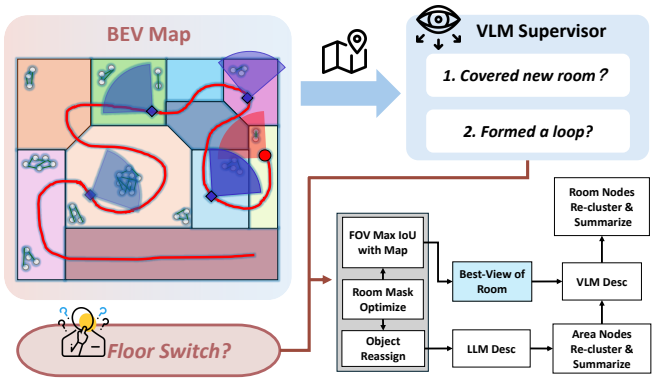

Fig. 5. **Event-Triggered Update.** Instead of fixed-frequency updates, our system monitors topological events. (Left) A BEV map tracks historical update points (Blue Wedges) and room transitions. (Right) When an update is triggered, the system selects representative observations to summarize the room's semantics and re-assigns objects to correct early segmentation errors.

When an update is triggered, the system first performs global room mask optimization and object re-assignment. It then initiates a bottom-up hierarchical generation. To instantiate Functional Areas (L_2), the system spatially clusters object nodes within each room. An LLM then processes their aggregated textual semantics to derive functional labels and summaries. Next, to construct the Room layer (L_1), the system selects a geometric Best-View frame. It identifies this frame by maximizing the intersection between the camera's field-of-view and the room's occupancy mask, while accounting for structural occlusions. We combine this optimal image with the generated L_2 summaries. Finally, a VLM synthesizes this multimodal context to produce high-level room descriptions. This event-driven approach allows the graph to evolve only when its semantic structure meaningfully changes, maintaining a consistent semantic forest structure without hindering the on-the-fly tracking of atomic objects.

D. Interpretable Closed-Loop Retrieval with Verification

A core limitation of existing semantic maps lies in how retrieval decisions are made. Vector databases often suffer from the attribute binding problem of logical negations. We fundamentally shift the retrieval paradigm from opaque recall-based embedding matching to an Interpretable Closed-Loop pipeline, adopting a physical implementation of the RAG workflow, as visualized in Fig. 6.

We first deploy a chain of specialized logical steps to decompose the complex human query. First, Constraint Decomposition breaks the raw instruction into atomic entity constraints, isolating target objects, reference landmarks, and spatial requirements. Next, Negation Extraction explicitly flags negative constraints, allowing the system to invert polarity during scoring. Finally, Intent Weighting interprets the user's semantic emphasis, assigning dynamic weights to attributes. For example, if the user emphasizes *the red one*, the retrieval module tends to prioritize color over location.

Instead of relying on a single similarity score, we formulate retrieval as constraint satisfaction process and employ a

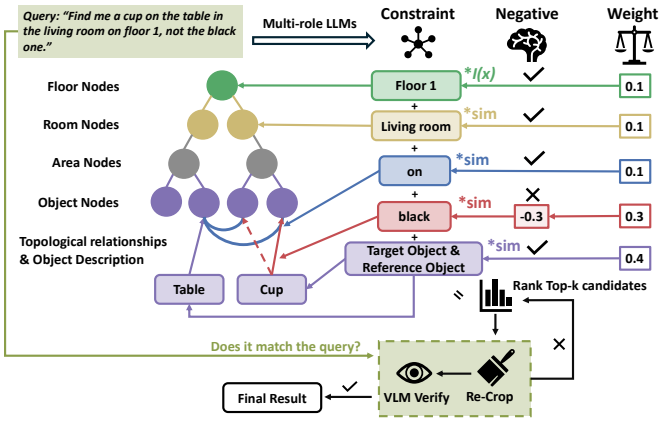

Fig. 6. **Interpretable Closed-Loop Retrieval.** Complex queries are decomposed by *Multi-role LLMs* into structural constraints, negative logic, and importance weights. The system performs hierarchical matching for a cumulative score. Top-ranked candidates undergo a final *VLM Verification* with re-cropping to ensure the result strictly aligns with the user's intent.

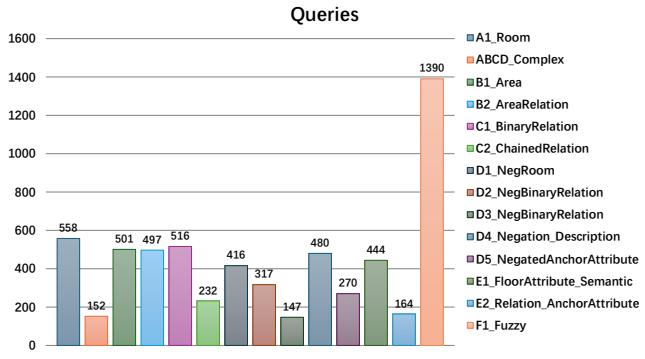

Fig. 7. **Distribution of Query Types.** The dataset encompasses a diverse range of complexities, spanning from basic spatial relations (A-C), descriptive queries (E), to challenging negations (D) and fuzzy descriptions (F).

hierarchical filtering strategy to rank candidates. The Floor ID serves as a binary *Hard Filter* ($H_{floor} \in \{0, 1\}$), immediately pruning the search space to the relevant level. Subsequently, all other parsed constraints function as *Soft Filters*. We calculate a composite relevance score $S(n)$ for each candidate node n by aggregating individual constraint scores. This process strictly adheres to the intent weights, formally defined as:

$$S(n) = H_{floor} \cdot \sum_{i=1}^{K} p_i \cdot w_i \cdot \text{Sim}(n, c_i) \tag{1}$$

where K is the number of constraints, w_i is the intent weight for the i-th constraint, and $\text{Sim}(n, c_i)$ denotes the semantic similarity between the node and the constraint. Crucially, $p_i \in \{1, -1\}$ represents the polarity indicator. This allows the system to positively score matching attributes and penalize nodes that satisfy negative constraints, ensuring the ranking aligns precisely with the user's specific intent.

Since similarity in feature space does not guarantee semantic correctness, an explicit verification step is required. Top-ranked candidates undergo a final *Visual Audit*. Here, a VLM validates the object against the specific query using the stored best-view image crop, eliminating feature-space misjudgments. Upon validation, the system outputs the precise 3D centroid. This allows seamless integration with downstream navigation tasks. Beyond retrieval, we demonstrate the lifelong potential of using natural language as object descriptions via a *Temporal Memory Fusion* cycle. Through a designed prompting strategy, the system can fuse the current interaction into the object's description while discarding outdated historical details and realize object-level temporal memory.

Furthermore, this retrieval architecture provides flexibility. The proposed strategies can be selectively composed to balance verification precision against computational latency. We provide a detailed quantitative analysis of these modules and their specific contributions in Table III. This closes the loop between language reasoning, visual evidence, and spatial memory.

IV. Experimental Evaluation

We design three types of experiments to comprehensively compare INHerit-SG with baselines: (i) Accuracy. We quantitatively compare INHerit-SG with recent open-vocabulary map representations in terms of retrieval accuracy on HM3DSem-SQR and real-world sequences (Sec. IV-B), (ii) Resource Usage. We analyze the memory usage of INHerit-SG compared to previous dense point-cloud representations (Sec. IV-C), and (iii) Ablation covering. We justify our design choices through a comprehensive ablation study covering hierarchy, timing, and verification modules (Sec. IV-D). Further, we design a multi-step navigation task in real-world environments based on validate the downstream effectiveness of INHerit-SG (Sec. IV-E).

A. Dataset and Baselines

Simulation Dataset. To evaluate whether semantic maps can support complex logical queries, we construct a dataset **HM3DSem-SQR** from HM3D-Sem [34], that stresses compositional reasoning rather than simple object recall. Unlike random sampling benchmarks, we employ human expert teleoperation to generate realistic exploration trajectories with synchronized sensor streams. Based on the trajectories, we manually constructed 36 trajectories (one per scene) and 6084 indexed instructions tailored to the characteristics of human commands and stress different requirements of a semantic map. Basic spatial relations (A-C) evaluate the need for a **Structured** multi-level topology. Negation queries (D) and descriptive queries (E) test whether the map is **Semantically Rich** enough to ground abstract concepts. descriptive queries (E), Ambiguous instructions (F) examine whether the system supports **Interpretable** reasoning beyond embedding similarity (Figure 7).

Realworld Dataset. We manually collected data from three real-world environments and designed 80 queries, evaluating the success rate through manual assessment in real scenes. The

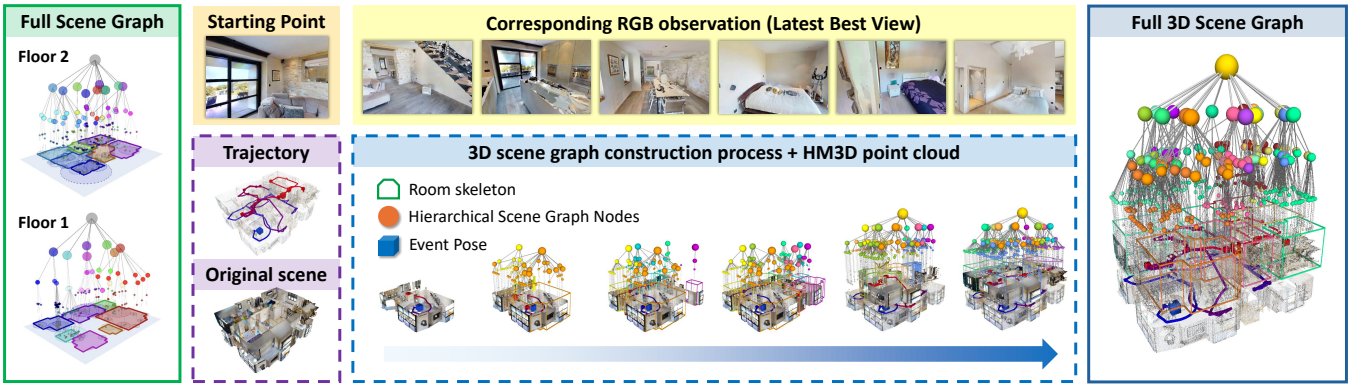

Fig. 8. **Qualitative Visualization INHerit-SG Construction.** We demonstrate the online generation of a hierarchical 3D scene graph on a multi-floor environment from the HM3D dataset. **(Left)** Distinct 2D scene graphs for illustration. **(Center)** The dynamic construction process. As the robot executes the trajectory, the system identifies key Event Poses (blue cubes) to trigger topological updates, incrementally expanding room skeletons (colorful outlines) and instantiating semantic nodes (spheres). **(Top)** Representative *Latest Best View* RGB observations. **(Right)** The final consolidated global 3D scene graph.

TABLE I
QUANTITATIVE COMPARISON ON CUSTOM DATASET (HM3DSEM-SQR) AND BENCHMARK (OPENLEX3D)

Method	HM3DSem-SQR Accuracy (%) ↑										Semantic Acc (%) ↑		Real-World Exp.		
	Within 1m					Within 0.5m									
	ABC	D	E	F	Avg	ABC	D	E	F	Avg	Random	Full Set	Simple	Complex	Avg
ConceptGraphs	22.84	14.79	21.54	20.22	19.95	21.62	14.30	21.38	18.99	19.03	–	–	27.3	44.4	35.0
ConceptGraphs(GPT)	13.48	13.38	9.05	13.38	12.98	–	–	–	–	–	–	–	–	–	–
Embodied-RAG	24.80	19.33	25.33	21.29	22.58	18.28	15.09	18.92	15.9	16.95	–	–	18.2	44.4	20.0
Embodied-RAG(GPT)	30.13	26.56	23.68	25.97	27.58	22.07	21.17	16.45	19.35	20.64	–	–	27.3	11.1	30.0
HOV-SG	27.0	31.6	34.7	28.5	29.40	20.32	23.07	25.33	22.01	21.94	–	–	–	–	–
DualMap	36.52	25.89	36.02	33.88	33.02	**30.78**	22.21	**31.58**	28.34	28.01	–	–	–	–	–
INHerit-SG (Ours)	**37.7**	**32.3**	**41.1**	**36.6**	**36.3**	30.1	**25.6**	30.9	**29.6**	**28.9**	**70.6**	**73.6**	**54.5**	**66.7**	**60.0**

camera trajectory was obtained from front-end SLAM system, while depth information was computed from a Livox LiDAR, providing the RGB-D stream and poses as input to our system. More details can be found in supplementary materials.

Baselines. We compare INHerit-SG against four state-of-the-art methods: **ConceptGraphs** [12] (flat, point-cloud based), **Embodied-RAG** [45] (open-loop retrieval), **HOV-SG** [44] (offline, hierarchical but offline), and **DualMap** [19] (SLAM-centric). All map construction are performed on a single RTX 4090 GPU with cloud-called GPT-4o.

B. Retrieval Accuracy

This experiment evaluates whether our representation and retrieval design improves reliability under complex semantic constraints. Since geometric precision is not the sole criterion in embodied tasks, we adopt two metrics: (i) **Geometric Accuracy**, measuring whether the retrieved object lies within a distance threshold of the ground truth, and (ii) **Semantic Accuracy**, assessing whether the object truly satisfies the instruction. To ensure fairness, the semantic metric is composed of two parts, including expert scoring over the full indexed query set, and a human study involving 120 participants who evaluated randomly sampled instructions.

Results (I) show that even under geometric-only evaluation, our method significantly outperforms all baselines at the 1.0m threshold. It maintains clear advantages at 0.5m on challenging queries such as negation and ambiguous semantics, and remain competitive on relatively easy cases. Despite not storing dense point clouds and operating under depth uncertainty, INHerit-SG remains highly competitive, demonstrating the benefit of its **Structured**. With human evaluation, semantic accuracy further improves. This demonstrates that once localization factors are excluded, the system intrinsically benefits from its **Semantically rich** representation and **Interpretable** retrieval aligned with human intent. On real-world data, INHerit-SG also demonstrates a clear advantage, highlighting its strong adaptability to noisy real environments. More details about human study and qualitative retrieval cases can be found in supplementary materials.

C. Resource Efficiency

A key design choice in INHerit-SG is replacing heavy point clouds with lightweight references, and treating the map as a knowledge base rather than a geometric container. Table II details the average storage consumption of all the simulation data.

We report two types of storage usage in Img. because

TABLE II
EFFICIENCY ANALYSIS BREAKDOWN

Method	Per-Object Node Storage (Avg)					Map Size
	Feat.	Img	Txt	PC	Node	(HM3D)
ConceptGraphs	4KB	21.33MB	4B	123.01KB	∼21.46MB	18.47GB
HOV-SG	22.3KB	–	–	28.3KB	∼94.2KB	1.79GB
DualMap	4KB	–	–	204.23KB	∼315.13KB	**87.4MB**
Ours	**21.1KB**	**405.0KB/-**	**155.8B**	**–**	**∼28.17KB**	47.5MB/**34.0MB**

TABLE III
ABLATION STUDIES ON COMPONENT CONTRIBUTION

Variant	**SR**	**Latency**
Full Model (INHerit-SG)	**74.0%**	**22.02 s**
Structural Ablations:		
1. w/o Functional Area Nodes (L_2)	71.7%	22.02 s
Retrieval & Semantic Ablations:		
2. w/o SAM3 (BBox only)	68.5%	20.36 s
3. w/o VLM Verification	65.4%	**11.75 s**

Note: Latency includes both tracking and mapping overhead.

our nodes store only lightweight reference pointers, while images are kept in a separate buffer. This separation means the reported map size does not depend directly on raw image storage. With straightforward image compression, our system offers substantial additional room for engineering optimization without altering the map structure itself. From Table II, most baselines rely heavily on dense point clouds, leading to bloated node sizes. As a result, our total map size is only 47.5MB, 34MB without images, achieving a sharp reduction compared to point-cloud-based methods.

D. Ablation Study

In order to shed light on the contributions of various key components in our approach, we present a comprehensive ablation study on a random sequence from HM3DSem-SQR in Table III. We evaluate both the Geometric Retrieval Accuracy and the average System Latency per query. Relying on cloud-based calls to large models, the measured latency is relatively high. With local deployment, it is reduced to approximately half. A detailed analysis is provided in the supplement.

Impact of Hierarchy and Architecture. Removing the *Functional Area Nodes* (Row 1) forces the system to search a larger, less structured graph, dropping accuracy by 2.3% and almost no increase in retrieval time. This again demonstrates the importance of a **Structured** multi-level topology for scalable reasoning.

Impact of Retrieval Components. Ablating *SAM3* (Row 2) and relying solely on bounding boxes significantly degrades accuracy (68.5%), showing that language descriptions alone are also insufficient, and must be grounded with precise visual perception to maintain a semantically rich representation.

Finally, removing *VLM Verification* (Row 3) results in the fastest system (11.75 s) but a drop in accuracy (65.4%). This indicates that storing visual references and performing verification substantially improves reliability, while also highlighting

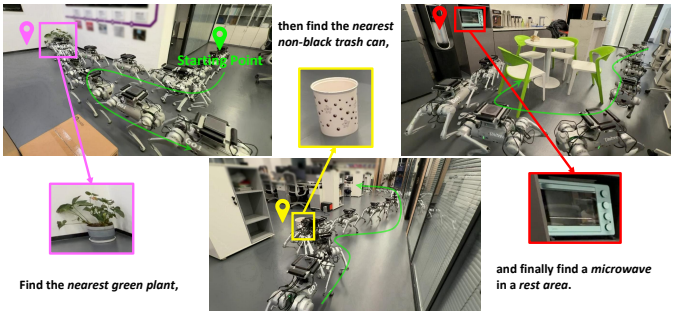

Fig. 9. **Real-World Navigation.** The robot successfully parses *Find the nearest green plant, then find the nearest non-black trash can, and finally find a microwave in a rest area.* retrieves the target, and navigates to it, validating metric accuracy.

that these components are not strictly required. The retrieval pipeline can be flexibly configured to trade off accuracy and latency, depending on task demands, demonstrating the modular and adaptable nature of our framework.

E. Qualitative Results: Downstream Integration

Finally, we demonstrate how the structured memory enables practical embodied behaviors. We use a Unitree Go1 robot connected to a cloud server to execute sequential tasks based on INHerit-SG retrievals. Integrated with ROS MoveBase, the system supports hierarchical planning on the Room layer before metric execution. This extends capabilities seen in SG-Nav [47]. Figure 9 demonstrates a successful "Find-and-Go" corresponding to the query in Figure 1.

V. CONCLUSION

In this work, we presented **INHerit-SG**, a framework that reframes semantic mapping as a structured, language-indexed knowledge base. By formulating the hierarchical scene graph as RAG-ready memory, we bridge geometric mapping with language-driven reasoning, replacing opaque embeddings with explicit, human-aligned descriptions that make spatial memory directly accessible for logical inference. We introduced a Event-Triggered map update mechanism that reorganizes topoligy only when meaningful semantic changes occur., enabling the graph to evolve incrementally as a long-term spatial memory. We further addressed the fragility of embedding-based retrieval by moving beyond similarity matching to a closed-loop verification process with logical parsing and visual auditing. Experiments confirm that INHerit-SG significantly suppresses misjudgments and effectively handles negation and chained relations where baseline methods fail.

Limitations and Future work. INHerit-SG currently assumes a relatively stable topology. While the event-triggered mechanism captures semantic transitions effectively, handling highly dynamic layouts or frequent object rearrangements remains challenging. Also, the retrieval pipeline relies on LLM/VLM reasoning, adding computational cost. Future work will seek more efficient yet interpretable alternatives and extend the framework to accommodate structural changes, and apply it to life-long scenarios and mobile manipulation tasks.

Acknowledgments

References

[1] Muhammad Qasim Ali, Saeejith Nair, Alexander Wong, Yuchen Cui, and Yuhao Chen. Graphpad: Inference-time 3d scene graph updates for embodied question answering, 2025. URL https://arxiv.org/abs/2506.01174.

[2] Abrar Anwar, John Welsh, Joydeep Biswas, Soha Pouya, and Yan Chang. Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation, 2024. URL https://arxiv.org/abs/2409.13682.

[3] Meghan Booker, Grayson Byrd, Bethany Kemp, Aurora Schmidt, and Corban Rivera. Embodiedrag: Dynamic 3d scene graph retrieval for efficient and scalable robot task planning, 2024. URL https://arxiv.org/abs/2410.23968.

[4] Carlos Campos, Richard Elvira, Juan J. Gomez Rodriguez, Jose M. M. Montiel, and Juan D. Tardos. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, December 2021. ISSN 1941-0468. doi: 10.1109/tro.2021.3075644. URL http://dx.doi.org/10.1109/TRO.2021.3075644.

[5] Yue Chang, Rufeng Chen, Zhaofan Zhang, Yi Chen, and Sihong Xie. Rag-3dsg: Enhancing 3d scene graphs with re-shot guided retrieval-augmented generation, 2026. URL https://arxiv.org/abs/2601.10168.

[6] Yinan Deng, Yufeng Yue, Jianyu Dou, Jingyu Zhao, Jiahui Wang, Yujie Tang, Yi Yang, and Mengyin Fu. Omnimap: A general mapping framework integrating optics, geometry, and semantics, 2025. URL https://arxiv.org/abs/2509.07500.

[7] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2025. URL https://arxiv.org/abs/2404.16130.

[8] Sourav Garg, Krishan Rana, Mehdi Hosseinzadeh, Lachlan Mares, Niko Sünderhauf, Feras Dayoub, and Ian Reid. Robohop: Segment-based topological map representation for open-world visual navigation, 2024. URL https://arxiv.org/abs/2405.05792.

[9] Luzhou Ge, Xiangyu Zhu, Zhuo Yang, and Xuesong Li. Dynamicgsg: Dynamic 3d gaussian scene graphs for environment adaptation. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, page 2232–2239. IEEE, October 2025. doi: 10.1109/iros60139.2025.11246569. URL http://dx.doi.org/10.1109/IROS60139.2025.11246569.

[10] Patrick Geneva, Kevin Eckenhoff, Woosik Lee, Yulin Yang, and Guoquan Huang. Openvins: A research platform for visual-inertial estimation. In *Proc. of the IEEE International Conference on Robotics and Automation*, Paris, France, 2020. URL https://github.com/rpng/open_vins.

[11] Nicolas Gorlo, Lukas Schmid, and Luca Carlone. Describe anything anywhere at any moment, 2025. URL https://arxiv.org/abs/2512.00565.

[12] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning, 2023. URL https://arxiv.org/abs/2309.16650.

[13] Tianjun Gu, Linfeng Li, Xuhong Wang, Chenghua Gong, Jingyu Gong, Zhizhong Zhang, Yuan Xie, Lizhuang Ma, and Xin Tan. Doraemon: Decentralized ontology-aware reliable agent with enhanced memory oriented navigation, 2025. URL https://arxiv.org/abs/2505.21969.

[14] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. 2024.

[15] Daniel Honerkamp, Martin Büchner, Fabien Despinoy, Tim Welschehold, and Abhinav Valada. Language-grounded dynamic scene graphs for interactive object search with mobile manipulation. *IEEE Robotics and Automation Letters*, 9(10):8298–8305, October 2024. ISSN 2377-3774. doi: 10.1109/lra.2024.3441495. URL http://dx.doi.org/10.1109/LRA.2024.3441495.

[16] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023.

[17] Nathan Hughes, Yun Chang, and Luca Carlone. Hydra: A real-time spatial perception system for 3d scene graph construction and optimization, 2022. URL https://arxiv.org/abs/2201.13360.

[18] Hanxiao Jiang, Binghao Huang, Ruihai Wu, Zhuoran Li, Shubham Garg, Hooshang Nayyeri, Shenlong Wang, and Yunzhu Li. Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation, 2024. URL https://arxiv.org/abs/2402.15487.

[19] Jiajun Jiang, Yiming Zhu, Zirui Wu, and Jie Song. Dualmap: Online open-vocabulary semantic mapping for natural language navigation in dynamic changing scenes. *IEEE Robotics and Automation Letters*, 10 (12):12612–12619, December 2025. ISSN 2377-3774. doi: 10.1109/lra.2025.3621942. URL http://dx.doi.org/10.1109/LRA.2025.3621942.

[20] Christina Kassab, Matías Mattamala, Sacha Morin, Martin Büchner, Abhinav Valada, Liam Paull, and Maurice Fallon. The bare necessities: Designing simple, effective open-vocabulary scene graphs, 2024. URL https://arxiv.org/abs/2412.01539.

[21] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *International Conference on Computer Vision (ICCV)*, 2023.

[22] Sebastian Koch, Narunas Vaskevicius, Mirco Colosi, Pedro Hermosilla, and Timo Ropinski. Open3dsg: Open-

vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships, 2024. URL https://arxiv.org/abs/2402.12259.

[23] Ryosuke Korekata, Quanting Xie, Yonatan Bisk, and Komei Sugiura. Affordance rag: Hierarchical multimodal retrieval with affordance-aware embodied memory for mobile manipulation, 2025. URL https://arxiv.org/abs/2512.18987.

[24] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

[25] Shiyang Lu, Haonan Chang, Eric Pu Jing, Abdeslam Boularias, and Kostas Bekris. Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data. In *7th Annual Conference on Robot Learning*, 2023.

[26] Dominic Maggio, Yun Chang, Nathan Hughes, Matthew Trang, Dan Griffith, Carlyn Dougherty, Eric Cristofalo, Lukas Schmid, and Luca Carlone. Clio: Real-time task-driven open-set 3d scene graphs, 2024. URL https://arxiv.org/abs/2404.13696.

[27] Yang Miao, Francis Engelmann, Olga Vysotska, Federico Tombari, Marc Pollefeys, and Dániel Béla Baráth. Scenegraphloc: Cross-modal coarse visual localization on 3d scene graphs, 2024. URL https://arxiv.org/abs/2404.00469.

[28] Siddarth Narasimhan, Matthew Lisondra, Haitong Wang, and Goldie Nejat. Splatsearch: Instance image goal navigation for mobile robots using 3d gaussian splatting and diffusion models, 2025. URL https://arxiv.org/abs/2511.12972.

[29] Phuc D. A. Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance, 2024. URL https://arxiv.org/abs/2312.10671.

[30] Phuoc Nguyen, Francesco Verdoja, and Ville Kyrki. React: Real-time efficient attribute clustering and transfer for updatable 3d scene graph, 2025. URL https://arxiv.org/abs/2503.03412.

[31] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[32] Xavi Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Ruslan Partsey, Jimmy Yang, Ruta Desai, Alexander William Clegg, Michal Hlavac, Tiffany Min, Theo Gervet, Vladimír Vondruš, Vincent-Pierre Berges, John Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A co-habitat for humans, avatars and robots, 2023.

[33] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4): 1004–1020, August 2018. ISSN 1941-0468. doi: 10.1109/tro.2018.2853729. URL http://dx.doi.org/10.1109/TRO.2018.2853729.

[34] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=-v4OuqNs5P.

[35] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning, 2023. URL https://arxiv.org/abs/2307.06135.

[36] Allen Z. Ren, Jaden Clark, Anushri Dixit, Masha Itkina, Anirudha Majumdar, and Dorsa Sadigh. Explore until confident: Efficient exploration for embodied question answering, 2024. URL https://arxiv.org/abs/2403.15941.

[37] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[38] Pranav Saxena and Jimmy Chiun. Zing-3d: Zero-shot incremental 3d scene graphs via vision-language models, 2025. URL https://arxiv.org/abs/2510.21069.

[39] Saumya Saxena, Blake Buchanan, Chris Paxton, Peiqi Liu, Bingqing Chen, Narunas Vaskevicius, Luigi Palmieri, Jonathan Francis, and Oliver Kroemer. Grapheqa: Using 3d semantic scene graphs for real-time embodied question answering, 2025. URL https://arxiv.org/abs/2412.14480.

[40] Lukas Schmid, Marcus Abate, Yun Chang, and Luca Carlone. Khronos: A unified approach for spatio-temporal metric-semantic slam in dynamic environments, 2024. URL https://arxiv.org/abs/2402.13817.

[41] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[42] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engel-

mann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[43] Yujie Tang, Meiling Wang, Yinan Deng, Zibo Zheng, Jingchuan Deng, and Yufeng Yue. Openin: Open-vocabulary instance-oriented navigation in dynamic domestic environments, 2025. URL https://arxiv.org/abs/2501.04279.

[44] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. In *Robotics: Science and Systems XX*, RSS2024. Robotics: Science and Systems Foundation, July 2024. doi: 10.15607/rss.2024.xx.077. URL http://dx.doi.org/10.15607/RSS.2024.XX.077.

[45] Quanting Xie, So Yeon Min, Pengliang Ji, Yue Yang, Tianyi Zhang, Kedi Xu, Aarav Bajaj, Ruslan Salakhutdinov, Matthew Johnson-Roberson, and Yonatan Bisk. Embodied-rag: General non-parametric embodied memory for retrieval and generation, 2025. URL https://arxiv.org/abs/2409.18313.

[46] Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F. Fouhey, and Joyce Chai. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent, 2023. URL https://arxiv.org/abs/2309.12311.

[47] Hang Yin, Xiuwei Xu, Zhenyu Wu, Jie Zhou, and Jiwen Lu. Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation, 2024. URL https://arxiv.org/abs/2410.08189.

[48] Chenyangguang Zhang, Alexandros Delitzas, Fangjinhua Wang, Ruida Zhang, Xiangyang Ji, Marc Pollefeys, and Francis Engelmann. Open-vocabulary functional 3d scene graphs for real-world indoor spaces, 2025. URL https://arxiv.org/abs/2503.19199.

[49] Lingfeng Zhang, Yuecheng Liu, Zhanguang Zhang, Matin Aghaei, Yaochen Hu, Hongjian Gu, Mohammad Ali Alomrani, David Gamaliel Arcos Bravo, Raika Karimi, Atia Hamidizadeh, Haoping Xu, Guowei Huang, Zhanpeng Zhang, Tongtong Cao, Weichao Qiu, Xingyue Quan, Jianye Hao, Yuzheng Zhuang, and Yingxue Zhang. Mem2ego: Empowering vision-language models with global-to-ego memory for long-horizon embodied navigation, 2025. URL https://arxiv.org/abs/2502.14254.

[50] Xiaolin Zhou, Tingyang Xiao, Liu Liu, Yucheng Wang, Maiyue Chen, Xinrui Meng, Xinjie Wang, Wei Feng, Wei Sui, and Zhizhong Su. Fsr-vln: Fast and slow reasoning for vision-language navigation with hierarchical multi-modal scene graph, 2025. URL https://arxiv.org/abs/2509.13733.

[51] Ziyu Zhu, Xilin Wang, Yixuan Li, Zhuofan Zhang, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Wei Liang, Qian Yu, Zhidong Deng, Siyuan Huang, and Qing Li. Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation, 2025. URL https://arxiv.org/abs/2507.04047.

[52] Xingxing Zuo, Pouya Samangouei, Yunwen Zhou, Yan Di, and Mingyang Li. Fmgs: Foundation model embedded 3d gaussian splatting for holistic 3d scene understanding, 2024. URL https://arxiv.org/abs/2401.01970.